DOCUMENT RESUME

ED 464 948 TM 033 879

AUTHOR Williamson, David M.; Johnson, Matthew S.; Sinharay, Sandip; Bejar, Isaac I.

TITLE Hierarchical IRT Examination of Isomorphic Equivalence of Complex Constructed Response Tasks.

INSTITUTION Educational Testing Service, Princeton, NJ.

PUB DATE 2002-04-00

NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Bayesian Statistics; *Constructed Response; Estimation (Mathematics); *Evaluation Methods; High Stakes Tests; *Item Response Theory; Markov Processes; Monte Carlo Methods; Test Construction

IDENTIFIERS *Calibration

ABSTRACT

This paper explores the application of a technique for hierarchical item response theory (IRT) calibration of complex constructed response tasks that has promise both as a calibration tool and as a means of evaluating the isomorphic equivalence of complex constructed response tasks. Isomorphic tasks are explicitly and rigorously designed to be highly similar in domain-relevant characteristics and evaluation standards. A related task model was used in which each item was modeled with a separate item response function, but the isomorphic tasks were related through a hierarchical model. The model was implemented in software that conducted Bayesian Markov Chain Monte Carlo (MCMC) estimation to estimate the joint posterior of all model parameters by integrating over the posterior distribution of model parameters given the data. The study analyzed operational data from a high-stakes assessment consisting of a number of complex constructed response tasks. The MCMC estimation procedure was conducted through 100,000 iterations. The item characteristic curves (ICCs) for the six isomorphic families were determined. In general, the families of isomorphic tasks showed considerable similarity in the item response functions for their respective members as well as for the family response function for the isomorphic set. Results suggest that efforts to construct complex constructed response tasks that are isomorphic equivalent tasks can range somewhat in their degree of success, with some being consistently equivalent, some being more variable, and others being largely consistent but with notable deviations. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

The following paper was prepared for presentation at the March 2002 AERA/NCME program and is the copyrighted material of Educational Testing Service. Limited permission to reproduce reports in their entirety in downloadable PDF format is hereby granted for individual professional reference. Requests for additional permission to reproduce or use all or portions of these papers may be made through the ETS Permissions website by completing the convenient form at: http://www.ets.org/legal/copyright/html

ED 464 948

Hierarchical IRT Examination of Isomorphic Equivalence of Complex Constructed Response Tasks

David M. Williamson

Matthew S. Johnson

Sandip Sinharay

Isaac I. Bejar

Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

D. Williamson

TM033879

BEST COPY AVAILABLE

Presented at the annual meeting of the American Educational Research Association
New Orleans, Louisiana
April, 2002

Hierarchical IRT Examination of Isomorphic Equivalence of Complex Constructed Response Tasks

When assessment design calls for the use of constructed response tasks the complexity of such tasks produces ripples of complexity throughout the assessment. This complexity impacts more than just the scoring, as issues such as the nature of the interface, availability of tutorials, and test development tools also impact the validity of an assessment for it's intended purpose (Bennett & Bejar, 1998). Resultant complication in assessment design is a concern in high-stakes assessment when the use of complex tasks may preclude the use of statistical techniques (e.g. equating) that are a mainstay of high-stakes multiple-choice testing. The field of educational measurement continues to seek methods to overcome this challenge for complex constructed response assessments. This paper explores the application of a technique for hierarchical item response theory (IRT) calibration of complex constructed response tasks which has promise as both a calibration tool and as a means of evaluating the isomorphic equivalence of complex constructed response tasks.

The use of complex constructed response tasks in high-stakes assessment presents multiple challenges not typically encountered in multiple-choice or low-stakes assessments. A central challenge stems from the extended period of time typically needed for examinees to complete complex domain tasks. It is not unusual for complex assessment tasks to require more than 30 minutes to complete and some high-stakes assessments use tasks that require two hours to complete. The extended time required for task completion prevents the administration of many such tasks in a single assessment

session. This can severely limit both the opportunity to pretest additional tasks and the feasibility of current equating techniques that rely on multiple items and imbedded equating blocks. In addition, when such tasks are computer-administered at secure testing sites any additional time allocated for additional tasks, such as pretesting, also incurs expense which must typically be passed on to the examinee. For complex tasks this additional expense may be non-trivial, particularly when a complex constructed response task in high-stakes testing can take up to two hours to complete.

A challenge for such an assessment environment, in which traditional equating cannot occur, is to maintain task security while simultaneously maintaining a uniform standard of evaluation. Particularly for assessments that provide continuous year-round testing the use of the same tasks for every administration permits even casual sharing of assessment experiences to provide unfair advantage to examinees who receive such knowledge in advance of their assessment. One possibility for addressing this challenge is through disciplined construction of task *isomorphs*; items that demand performance of the same domain tasks, use identical features in scoring, have highly similar statistical performance, and measure the same knowledge and skills, but appear to be substantially different items. By carefully constructing a number of different tasks so that they maintain a highly consistent set of required domain activities and substantially varying the surface features of the task (e.g. context, incidental details, etc.) the intent is to create items which perform in a substantially similar fashion but present to the casual observer as independent items. The intent of such isomorphic tasks is to serve as a pool of interchangeable items which can be drawn at random to create alternate test forms for examinees. For assessments that use automated scoring to evaluate the work products

from complex constructed response tasks the production of isomorphic items is facilitated by the requirement that each task be scorable using an identical automated scoring algorithm. The facilitation from automated scoring algorithms rests with the algorithms explicitly defined criteria and pre-programmed expectations about the nature of the work product produced from such tasks, thus contributing to the rigor with which isomorphic tasks are created.

The development and implementation of isomorphic tasks in high-stakes assessment begs the question of how to properly consider this isomorphic equivalence when calibrating complex tasks. Given that isomorphic tasks are explicitly and rigorously designed to be highly similar in domain-relevant characteristics and evaluation standards it may be expected that isomorphic tasks would have a considerable degree of similarity in both content and statistical performance. This expectation leads to a fundamental question of how to best model such isomorphic siblings in operational measurement. To that question there are at least three possible techniques for IRT calibration of such items.

Unrelated Task Model

The most conservative approach for calibration of isomorphic tasks is to treat them as completely independent despite the fact that they share a strong fundamental similarity. This unrelated task model is given by

$$P_j(\theta) = \frac{1}{1+\exp\{a_j(\beta_j - \theta)\}} \qquad (1)$$

where *j* indicates the particular task in question. Since the model ignores the relationship between isomorphs the model is overly conservative, with use of these item response functions resulting in an unnecessarily large standard error for θ estimates.

Identical Task Model

A more liberal approach to calibration of item isomorphs is to consider them as having identical item response functions (Hombo & Dresher, 2001). This model is given by

$$P_j(\theta) = \frac{1}{1+\exp\{a_{I(j)}(\beta_{I(j)} - \theta)\}} \qquad (2)$$

where *I(j)* indicates the isomorph set of which task *j* is a member. Since the identical isomorph model ignores all variation between isomorphic tasks it results in inappropriately small standard errors for θ estimates, reflecting overconfidence about the ability of the examinee.

Related Task Model

A third alternative, utilized in the analyses for this paper, is to use a related task model in which each item is modeled with a separate item response function, but the isomorphic tasks are related through a hierarchical model (Glas & van der Linden, 2001).

$$P_i(\theta_i) \equiv \Pr\{X_{ij} = 1 | \theta_i\} = \frac{1}{1+\exp\{\alpha_i(\beta_i - \theta_i)\}} \quad (3)$$

where

$\theta \sim N(\mu, \sigma^2)$

$\alpha_i \equiv \log\{\alpha_i\}$

$(\alpha_i, \beta_i)' \sim N_2(\lambda_{I(i)}, T_{I(i)})$

and where *i* indicates the examinee in question. This model appropriately accounts for sources of variation in responses: The responses of two individuals answering the same isomorph are correlated. An additional advantage of this approach is that calibration of the isomorphic family and use of a family response function requires substantially fewer observations for each isomorph than calibration of each isomorph individually.

This model is implemented in software (Johnson & Sinharay, April, 2002) that conducts Bayesian Markov Chain Monte Carlo (MCMC) estimation to estimate the joint posterior of all model parameters by integrating over the posterior distribution of model parameters given the data. The Monte Carlo integration draws samples from the required distribution and then forms sample distributions to approximate expectations. MCMC draws these samples by running a Markov chain through many iterations. As such, MCMC estimation is basically Monte Carlo integration using Markov chains; discrete time stochastic processes such that the distribution of X_t (X at time t) depends only on X_{t-1}

and is independent of all values X_{t-1} to X_{t-n}. Mathematically, this is represented as (Gilks, Richardson, & Spiegelhalter, 1996, p. 45):

$$P[X_t \in A \mid X_0, X_1, \ldots, X_{t-1}] = P[X_t \in A \mid X_{t-1}] \quad (4)$$

for any set A, where $P[.|.]$ denotes a conditional probability. For the related siblings model MCMC estimates the posterior distribution by drawing from the conditional posterior distribution of each model parameter. Item parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are drawn from their respective conditional distributions as described in Patz and Junker (1999). Conditional on the item parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the item family mean vector $\boldsymbol{\lambda}$ and the covariance matrix $\mathbf{T}$ are independent of θ and the observed data $\mathbf{X}$.

This study applies the related task model as implemented in software by Johnson & Sinharay (April, 2002) to operational data from a high-stakes complex constructed response assessment using isomorphic items. This application examines the correspondence between item characteristic curves for individual isomorphic items and the item characteristic curves for the isomorphic family. Use of such a hierarchical calibration both reflects the rigorous design of the tasks for isomorphic equivalence and allows the examination of the similarity of item response functions to common isomorphic family response function that could be used in the operational assessment.

Method

Data

This study analyzed operational data from a high-stakes assessment consisting of a number of complex constructed response tasks, each of which are scored on a 3-point polytomous scale. Each administration of the assessment consists of six tasks; one from

each of six distinct task domains. Within each domain are a family of tasks constructed to be isomorphic equivalent tasks (i.e. demand performance of the same domain tasks, use identical features in scoring, have highly similar statistical performance, and measure the same knowledge and skills, but by virtue of substantial changes to surface features appear to be substantially different items). For any particular administration the task is drawn at random from the task family for the domain in question. That is, there are six pools of isomorphic tasks and for any given examinee one task is drawn at random from each of the six pools to construct the examinee's assessment. A breakdown of the sample size by isomorphic task pool is provided as Table 1.

Procedure

Analysis of the data was conducted with software (Johnson & Sinharay, April, 2002) that calibrates items using a hierarchical model (Glas & van der Linden, 2001) as described above. The current version of the software is designed for dichotomous cases only, with the extension to polytomous cases under current development. Therefore, the polytomous data was transformed into dichotomous data by collapsing the two lowest score categories into a single response category. The model applied prior distributions for the item family mean vectors that assumes the elements are independent and

$\lambda_a \sim N(0,100^2)$

$\lambda_b \sim N(0,100^2)$.

The MCMC estimation procedure was conducted through 100,000 iterations, with the first 10,000 iterations treated as a burn-in period and therefore not included in the determination of the posterior distributions of the parameters. The remaining 90,000

iterations were thinned by selecting every 9th iteration for inclusion in the final data set determining the posterior distribution of the parameters. This resulted in a final data set consisting of 10,000 draws for the distribution of each parameter. The item characteristic curves (ICC) were produced using the median value of the distribution for each parameter. The root-mean-square-error (RMSE) was computed for the ICCs for each isomorphic group, using the group calibration as the ICC for comparison of the item ICCs in the computation. The RMSE is given by

$$RMSE = \sqrt{\frac{\sum_{t=-3.0}^{3.0} (p_a - p_B)^2}{n_t}} \qquad (5)$$

where p_a indicates the item ICC probability of responding correctly at ability t, p_B indicates the family ICC probability of responding correctly at ability t, and n_t is the number of theta values considered (in this case using the values between –3.0 and 3.0 in intervals of .1, so n_t=61).

Results

The ICCs for the six isomorphic families are provided as Figure 1. The greatest degree of variation in the task ICCs from the ICC for the family as a whole is in family B1 while the least variation is observed in family B8. The ICCs for family B4 are similar with one notable exception in task B430. In general, the families of isomorphic tasks showed considerable similarity in the item response functions for their respective members as well as for the family response function for the isomorphic set.

The plot of RMSE for the tasks in each isomorphic family are provided as Figure 2. The notable peak in RMSE for item three of isomorphic family B4 reflects the extreme case visible in the ICC for B4.

Discussion

These results suggest that efforts to construct complex constructed response tasks that are isomorphic equivalent tasks can range somewhat in their degree of success, with some being consistently equivalent (e.g. family B8), some being more variable (e.g. family B1), and others being largely consistent but with notable deviations (e.g. family B4).

The range of RMSEs computed from these isomorphic items is similar to the range of RMSEs obtained from a study (Rizavi, Way, Davey, & Herbert, April, 2002) in which the same subset of items from Verbal and Quantitative sections of a high-stakes admissions test were recalibrated through eight administrations and the variation in item parameters evaluated. If variations in ICCs for isomorphic constructed response tasks are consistently similar to variations obtained from recalibration of an identical multiple-choice item then the goal of creating isomorphic constructed response tasks with highly similar statistical performance has been largely met.

Despite the similarity of RMSEs between calibrations of complex constructed response isomorphs and recalibrations of the same subset of items on a high-stakes admissions test, there remain issues to be studied regarding the impact of such variation, both for isomorphic variants and for variation due to multiple-choice item calibration. Specifically, it will be important to establish the degree of variation in θ estimates as a

result of such variation and, for case of assessments designed for classification (i.e. licensure, placement, etc.) the impact on such classification decisions. This area of investigation is ripe for additional research, both for multiple-choice calibration variations as well as variations in the isomorphic equivalence of complex constructed response items.

Researchers in the field have recognized the importance of these issues and have already begun to address them. Dresher & Hombo (2001), for example, investigated the impact of simulated parameter variation on ability estimation and concluded that ability estimation, for both individuals and grouped score reporting, was largely robust to variation in parameter estimates. The impact of item parameter variation on ability estimates has also been addressed by by Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta (in press). As a result of these investigations the field is developing a clearer perspective on the feasibility and appropriateness of isomorphic item modeling as a means of addressing difficulties inherent in the use of highly complex constructed response tasks in high-stakes assessment.

References

Bejar, I. i., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (in press). *A feasibility study of on-the-fly adaptive testing* (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Bennett, R. E. & Bejar I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17(4)*, 9-17.

Dresher A, & Hombo, C. (2001). *A simulation study of the impact of automatic item generation on item and ability parameter estimation*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov chain monte carlo in practice*. New York: Chapman & Hall.

Glas, C. A. W. & van der Linden, W. J. (2001). *Modeling variability in item parameters in CAT*. Paper presented at the North American Psychometric Society Meeting, King of Prussia, PA.

Hombo, C. & Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Johnson, M. S. & Sinharay, S. (April, 2002). *Hierarchical approaches to item model calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Patz, R. & Junker, B. (1999). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342-366.

Rizavi, S., Way, W. D., Davey, T., & Herbert, E. (April, 2002) *Tolerable variation in item parameter estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Author Note

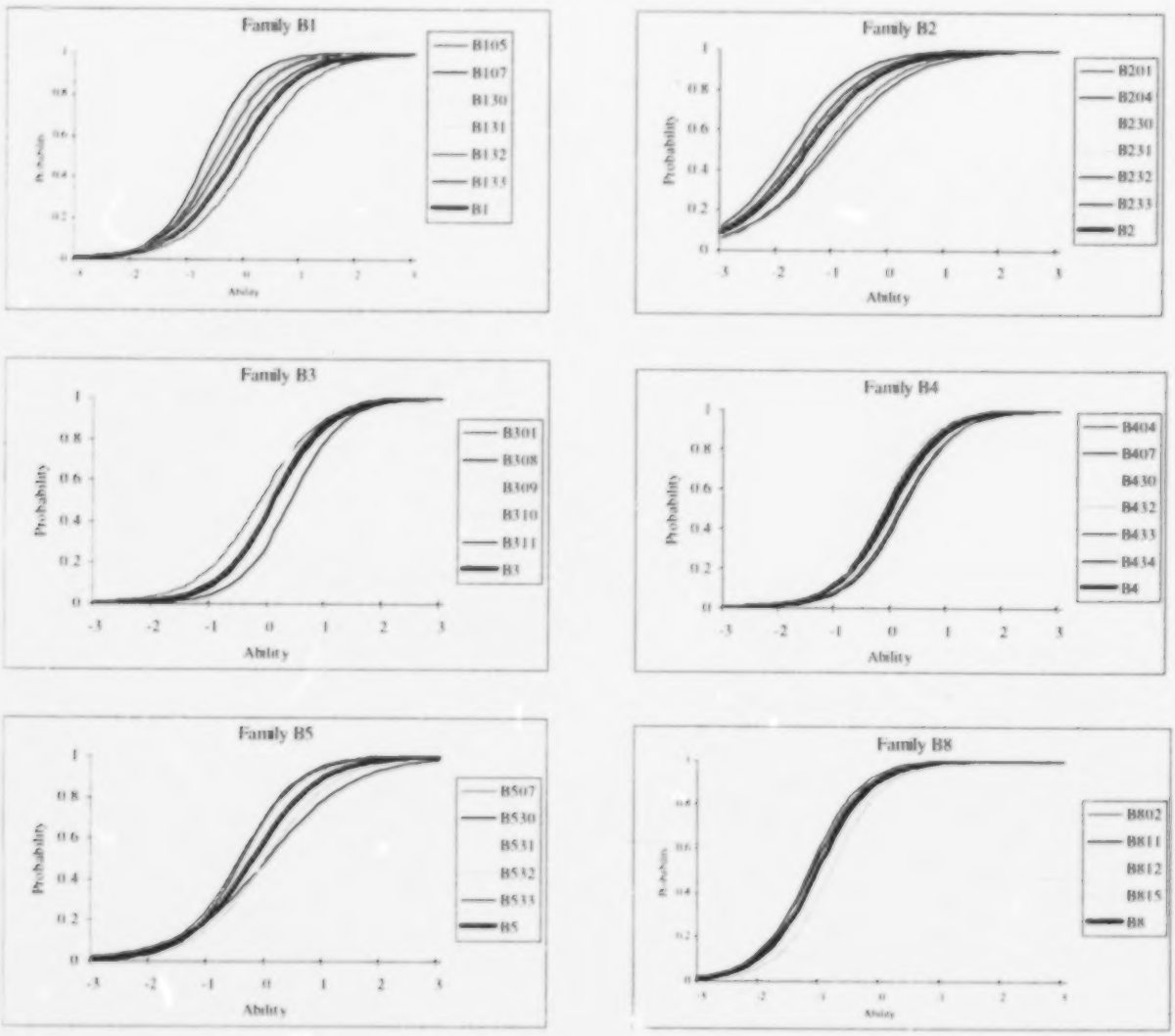
The authors thank Chris Chiu for conducting some of the preliminary analytical work preparing the data for this study.

Table 1
Sample Sizes for Isomorphic Sets

Task Set	Sample Size
B1	572
B2	575
B3	571
B4	572
B5	518
B8	511

Figure 1

Hierarchical IRT Model Item Characteristic Curves by Isomorphic Family



BEST COPY AVAILABLE

Figure 2

Root Mean Squared Errors Between Item Characteristic Curves and Isomorphic Family Characteristic Curves